دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده مهندسی کامپیوتر

داده کاوی

(بهار ۱۴۰۱)

تمرین دوم

محمد چوپان ۹۸۳۱۱۲۵

بخش تئوری :

سوال اول :

سوال اول

یکی از مباحثی که در درخت تصمیم مطرح میشود هرس درخت [1] برای جلوگیری از بیش برازش است. توضیح دهید چرا نمیتوان از مجموعه داده جدا برای هرس درخت استفاده کرد؟ منظور این است که داده‌هایی که برای هرس استفاده میشوند با مجموعه داده‌ای که برای ساخت درخت استفاده میشود یکسان نباشند.

پاسخ:

هدف از هرس درخت این است که برگهایی که برای داده های آموزش دیده نمی‌شوند به احتمال حذف شوند تا از بیش برازش در داده جلوگیری شود. بنابر هدف اصلی هرس درخت این است که از overfitting در داده ای آزمایش جلوگیری کند. اما اگر از مجموعه داده‌ای که برای هرس درخت استفاده میشود داده های جدایی باشند این باعث خواهد شد که درخت با داده های آزمایشی به درستی کار نکند و احتمال overfitting همچنان باقی می ماند .بنابراین برای هرس باید از همان داده های آموزش استفاده شوند. تا درخت بتوان به صورت صحیحی کار کند.

سوال دوم :

سوال دوم

با توجه به مطالب تدریس شده در کلاس، برای داده‌های زیر یک درخت تصمیم درست کنید. (ذکر تمام مراحل و توضیح آنها لازم است)

آیا به مهمانی دعوت میشود؟	وزن	قد	رنگ لباس
خیر	لاغر	۱۷۰	قرمز
بله	چاق	۱۶۲	آبی
خیر	چاق	۱۶۵	سبز
بله	لاغر	۱۷۲	سبز
بله	لاغر	۱۶۰	آبی

پاسخ:

سوال ۲ :

۱ ابتدا باید آنتروپی داده های ورودی را حساب کنیم

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.97$$

$$Info_A(D) = \sum_{i=1}^{v} \frac{|D_i|}{|D|} \times Info(D_i)$$

Feature A

$$Info_{\text{رنگ}}(D) = \frac{1}{5}I(1) + \frac{2}{5}I(1) + \frac{2}{5}I(1) = 0.4$$

$$Info_{\text{وزن}}(D) = \frac{3}{5}I(2,1) + \frac{2}{5}I(1,1) = 0.95$$

برای دیتگی قدیمی توانیم به ۳ بازه دسته بندی کنیم

فرض کنیم ۱ متفاوت ② و ④ برای قد در نظر بگیریم

قد ≤ ۱۶۵ — ⑤ — yes

قد > ۱۶۵ — ③ — no

$$Info_{\text{قد}}(D) = \frac{4}{5}I(2,2) + \frac{1}{5}I(1,0) = 0.8$$

$$Info_{\text{قد}} = \frac{3}{5}I(2,1) + \frac{2}{5}I(2,0) = 0.55$$

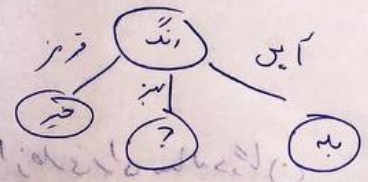
پس بهترین میزان ① مناسب ترین ویژگی برای جداسازی

$$Gain_{\text{وزن}} = 0.97 - 0.95 = 0.02$$

$$Gain_{\text{رنگ}} = 0.97 - 0.4 = 0.57$$

$$Gain_{\text{قد}} = 0.97 - 0.55 = 0.42$$

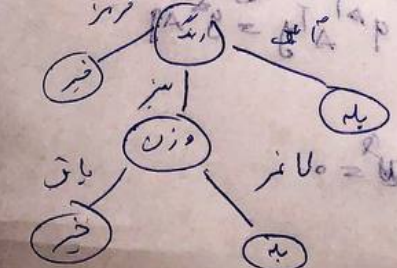
پس رنگ به ریشه ی درخت میرود.



حال برای رنگ سبز $Info(D) = 1$

$$Gain_{\text{وزن}} = 1 - 0.95 = 0.05$$

$$Gain_{\text{قد}} = 1 - 0.8 = 0.2$$



وزن ریشه جدید

نیاز به قد ندارد.

سوال سوم :

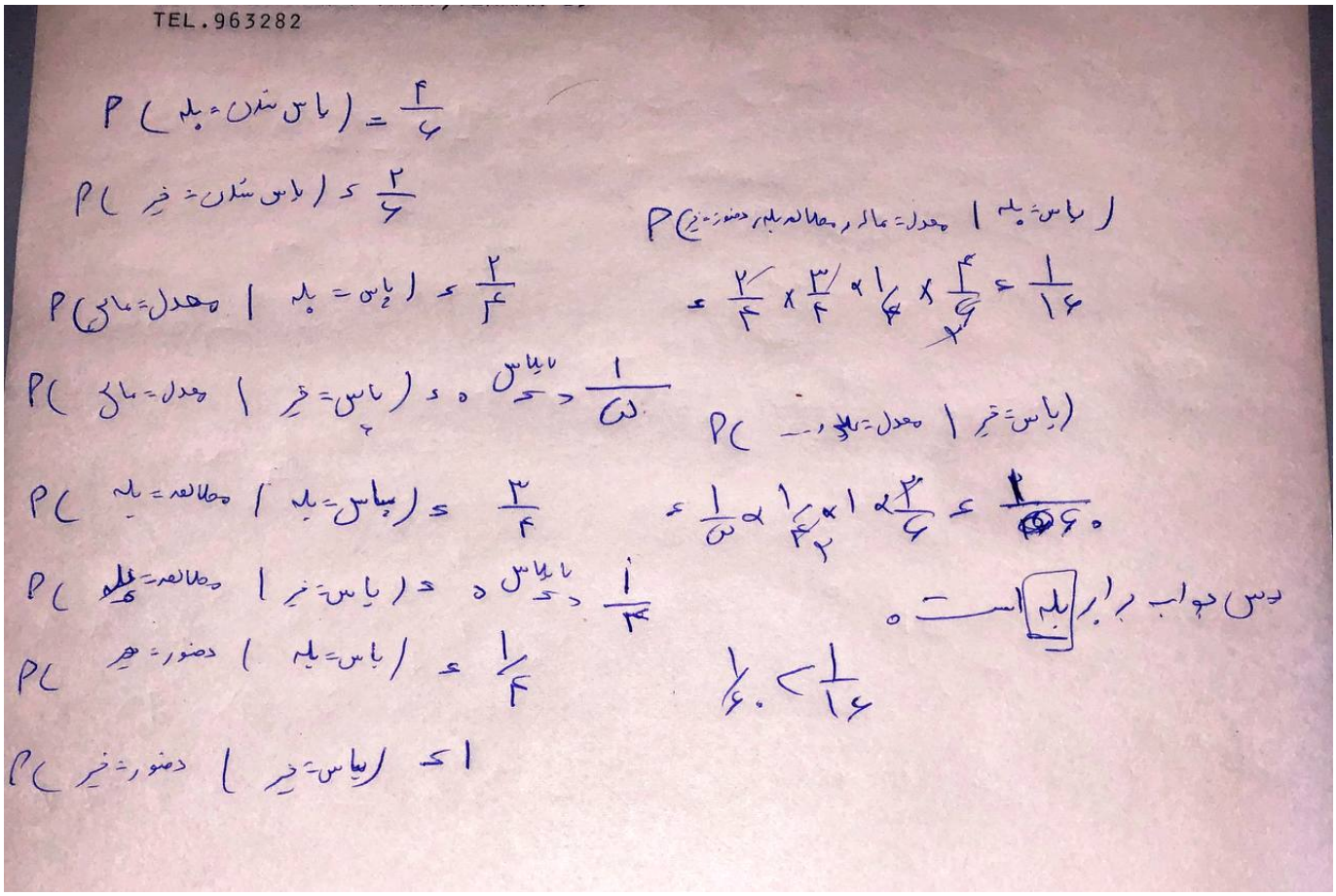
سوال سوم

در جدول داده شده زیر با استفاده از قانون بیز برچسب داده زیر را به دست آورید. در صورت صفر شدن احتمال از هموارسازی لاپلاس[۲] استفاده کنید.

(معدل = عالی ، مطالعه = بله ، حضور = خیر)

پاس شدن	حضور در کلاس‌ها	مطالعه برای امتحان	معدل
خیر	خیر	خیر	ضعیف
بله	بله	بله	ضعیف
خیر	خیر	خیر	متوسط
بله	بله	بله	متوسط
بله	خیر	خیر	عالی
بله	بله	بله	عالی

پاسخ:



سوال چهارم :

سوال چهارم

همانطور که میدانیم یکی از معیارها برای ارزیابی مدل‌های یادگیری نظارت شده صحت [۳] است. اما این معیار در برخی موارد ممکن است معیار مناسبی برای ارزیابی نباشد. موقعیت‌هایی که این معیار برای ارزیابی به خوبی عمل نمیکند را توضیح دهید.

پاسخ :

چند مورد است که این معیار به خوبی عمل نمی‌کند:

۱. داده های پرت : وجود داده های پرت و یا نویز ها باعث می‌شود که صحت مدل ما کم شود . داده های پرت ممکن است منجر به اشتباهات طبقه بندی شود و صحت را تحت تاثیر قرار دهند.

۲. خطاهای نوعی: در برخی مسائل اهمیت خطاهای نوعی مانند False Positive , False negative ممکن است متفاوت باشد مثلا تشخیص یک بیمار مریض به عنوان سالم بسیار مهم است. این حالت ها نیز باعث می‌شوند تا نتایج دقیقی ارائه نشوند و عملکرد واقعی مدل این نباشد.

۳. دسته بندی نامتوازن: زمانی که توزیع داده ها نا مناسب باشد یعنی داده های یک کلاس خیلی کمتر از دیگری باشد . معیار صحت تنها از تعداد درست دسته بندی شده ها را اعلام میکند. ممکن است مدل به طور غیر مناسب برچسب کلاس اقلیت را پیش بینی کند و صحت بالا داشته باشد در حالی که به طور کلی برای تشخیص الگوها و روابط در داده ها به مشکل برخورد کند.

سوال پنجم :

سوال پنجم

فرض کنید که برای انتخاب پارامتر α در مدل از روش 10 fold cross validation استفاده کرده‌ایم. بهترین روش برای انتخاب مدل نهایی و تخمین ارور کدام است؟

پاسخ :

روش ها ی زیادی برای این کار وجود دارند و که ما دو تا از آن ها را می نویسیم:

- روش میانگین دقت : در این روش میانگین دقت حاصل از اجرای ۱۰ فولد مختلف برای داده ها را محاسبه میکنید. سپس مدلی را اتخاب میکتید که داراب بیشترین میانگین قیمت باشد و این روش به شما ایده ی کلی از عملکرد مدل در داده های جدید می دهد.
- روش انتخاب بر اساس خطا (ارور): در این روش میانگین ارور یا همان خطا را برای هر فولد محاسبه می کنید. سپس مدلی را انتخاب میکنید که دارای کمترین میانگین ارور باشد . در این روش به شما ایده ی کلی از کارایی مدل در پیش بینی داده های جدید می دهد.

در هر دو روش استفاده از هر ۱۰ تا فولد این ویژگی را به ما میدهد مه اعتماد جامع تر و قابل اعتماد تری داشته باشیمو برای اینکه مطمئن تر شویم نیز می‌توانیم از معیار هایی مانند واریانس و انحراف معیار نیز استفاده کنیم. که مدل را ارزیابی کنیم.

سوال ششم :

سوال ششم

در الگوریتم boosting اگر هر کدام از موارد زیر رخ دهد ما یادگیری را متوقف میکنیم؟ برای پاسخ‌های خود دلیل بیاورید.

- میزان خطای طبقه‌بندی‌کننده ترکیبی در داده‌های آموزشی اصلی ۰ شود.
- میزان خطای طبقه‌بندی‌کننده ضعیف[۴] فعلی روی داده‌های تمرین وزن‌دار[۵] ۰ است.

پاسخ :

الف : در حالت اول بله در الگوریتم boosting اگر میزان خطا طبقه بندی ترکیبی در داده های آموزشی اصلی به صفر برسد یعنی ترکیب طبقه بندی های ضعیف به گونه ای توانسته است که همه نمونه های داده های آموزشی را به درستی طبقه بندی کند در این حالت مدل دیگر نیازی به یادگیری بیشتر ندارد زیرا قادر به دقیق تشخیص دادن داده های آموزشی است. همچنین متوقف کرد این کار باعث می‌شود که دیگر overfitting دیگر رخ ندهد. وقتی که خطا به صفر میرسد به احتمال خیلی بالا مدل در حال overfitting است به ادامه دادن این فرایند در یادگیری ممکن است منجر به افزایش خطا در داده های تست شود . بنابراین اگر چه همه داده ها را دست تشخیص می دهد اما بهتر است که از آن جلوگیری کرد.

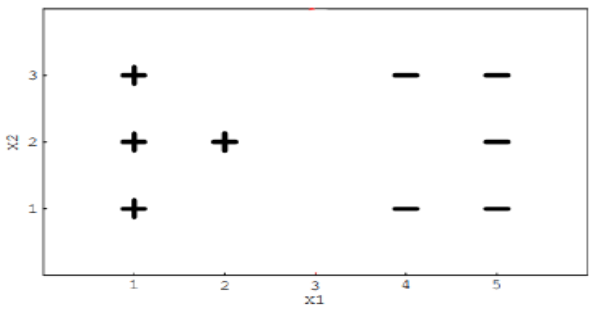
ب: خیر نیازی به متوقف کردن آموزش در صورت رسیدن به خطای ۰ بر روی داده های وزن دار در الگوریتم boosting نیست. به دلیل یاینکه مدل باید الگو های پیچیده را پیدا کند و این الگو ها را تعمیم دهد. رسیدن به خطای صفر در این داده ها میتواند نشانه بیش برازش باشد اما این موضوع تنهایی نمی‌تواند بر این دلالت کند که مدل در حال بیش برازش است. امکان دارد که ادامه آموزش به مدل کمک کند و الگو های جدید را پیدا کند. پس بهتر است ادامه یابد تا مطمئن شویم که بیش برازش رخ نداده و الگو های جدید نیز پیدا شوند و تعمیم پیدا کنند.

سوال هفتم :

سوال هفتم

فرض کنید برای داده‌های زیر از طبقه‌بندی کننده SVM خطی بدون کرنل استفاده میکنیم و پارامتر C در این طبقه بندی کننده بسیار بزرگ در نظر گرفته شده است. (اگر در مورد این پارامتر اطلاعی ندارید این لینک را مطالعه کنید).

الف) خطی که SVM گفته شده با استفاده از آن داده ها را دسته میکند را رسم کنید و علت انتخاب این خط را توضیح دهید.



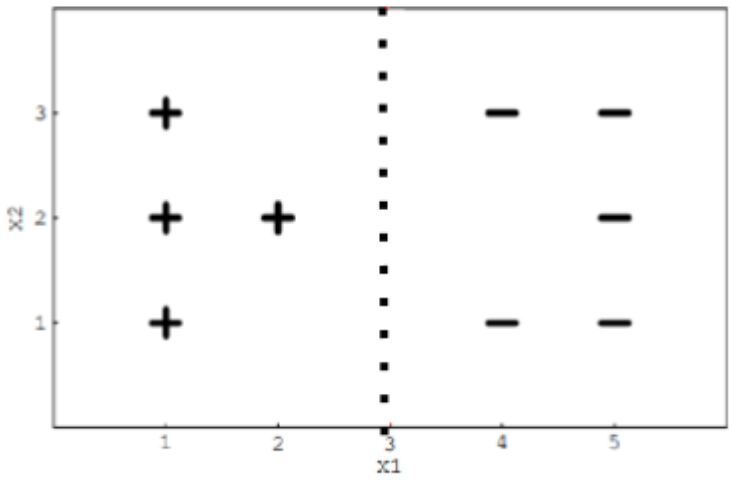
ب) در شکل بالا نقاطی را انتخاب کنید که حذف آنها باعث میشود خطی که SVM داده ها را جدا میکند متفاوت از حالت **(الف)** شود. دلیل انتخاب این نقاط را توضیح دهید.

پاسخ :

الف :

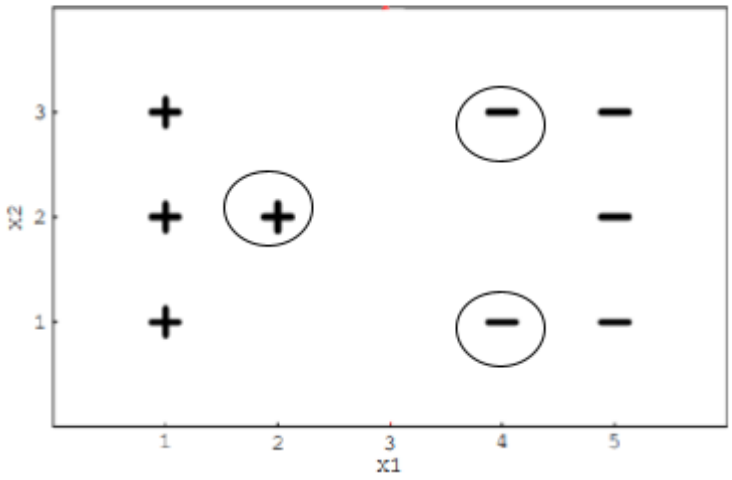
در این نوع دسته بندی باید بیشترین فاصله با داده های مرزی وجود داشته باشد تا زمانی که یک داده جدید وارد می شود به داده درست وارد شود.بنابر خطی که از وسط این دو می گذرد بهترین خط است. و با بزرگ گرفتن پارامتر C باعث می شود که margin ما کوچکتر شود و خط جدا کننده وسط باشد.

شده با استفاده از ان داده ها را دسته میکند را رسم کنید و علت انتخاب این خط .



ب: نقاطی که باعث می شود خط ما وسط بیافتند نقاط مرزی هستند که در شکل زیر مشخص است چرا که فاصله با نقاط مرزی محاسبه می شود حال اگر ما این ها را برداریم خط جدا کننده میتواند جدا شود.

ه با استفاده از ان داده ها را دسته میکند را رسم کنید و علت انتخاب ا



سوال هشتم :

سوال هشتم

صحیح یا غلط بودن موارد زیر را با دلیل مشخص کنید :

الف) الگوریتم بیز ساده نمیتواند وابستگی بین متغیرها را مشخص کند.

ب) هنگامی که یک درخت تصمیم به سمت یک درخت پر پیش میرود احتمال اینکه نویز را هم پوشش دهد بیشتر میشود.

ج) در روش k نزدیک ترین همسایه[۶] اگر k=1 الگوریتم نسبت به داده‌های نویز مقاوم تر از حالتی است که k=5 در نظر گرفته شود.

پاسخ :

الف :

درست است. به دلیل اینکه در این الگوریتم احتمال ها کاملا مستقل از هم هستند و بدون توجه به تخمین پارامتر های دیگر می پردازد. یعنی بیز ساده فرض مستقلیت شرطی را در متغیر ها اعمال می کند و از وابستگی متغیر ها به هم چشم پوشی میکند.

ب:

نادرست است. اینکه نویز ها را پوشش بدهد بیشتر یه این که درخت ما بیش برازش یشود یا نه بستگی دارد نه به اینکه به سمت یک پر پیش بریم یعنی امکان دارد که نویز ها باعث کاهش دقت عملکرد و دقت درخت روی داده های ما شوند و این با پیشروی اتفاق نمیافتد همچنین هرچه به سمت پر ها می رویم تاثیر تصمیم گیری بر روی داده های کوچکتر کمتر میشود و احتمال اینکه پوشش بدهد کمتر است.

ج :

نادرست است. در این الگوریتم هرچه که k کمتر باشد الگوریتم حساسیت بیشتری نسبت به نویز ها دارد و در نتیجه مقاومت کمتری در برابر نویز ها دارد و یعنی در این حالت الگوریتم به اندازه کافی داده های مشابه هم را بررسی نمیکند و ممکن است داده های نویز بیش از حد احساس شوند. و اگر که k زیاد شود الگوریتم داده های نزدیک تر را بیشتر بررسی میکند و نویز ها تاثیر کمتری را بر روی نتیجه نهایی دارند. این به معنی است که افزایش مقاومت الگوریتم در برابر نویز است.

سوال نهم :

سوال نهم

فرض کنید در حال طراحی یک سیستم برای تشخیص خستگی راننده در اتومبیل هستید. بسیار مهم است که مدل شما خستگی را تشخیص دهد تا از هر گونه حادثه ای جلوگیری شود. کدام یک از معیارهای زیر بهترین معیار برای ارزیابی هست : Accuracy, Precision, Recall, Loss Value دلیل انتخاب خود را شرح دهید.

پاسخ :

به نظر من پاسخ درست recall است به دلیل اینکه اینجا هدف اصلی ما تشخیص خستگی راننده است پس باید از یکی از معیار های ارزیابی استفاده کنیم همچنین میتوانیم از چند معیار مانند precision و accuracy هم

برای دقیق تر شدن استفاده کنیم. اما به دلیل حالت های FP یعنی نادرست مثبت به ما بدهند بهتر از این است که درست مثبت ها را به ما ندهند پس معیاری مانند recall هزینه کمتری برای تصادف دارد.

سوال دهم:

سوال دهم

علاوه بر شاخص آنتروپی برای ساخت درخت تصمیم، شاخص دیگری نیز وجود دارد که میتوان به جای آنتروپی از آن برای ساخت درخت استفاده کرد. این شاخص را معرفی کنید و بگویید تفاوت آن با آنتروپی چیست؟ بالا یا پایین بودن این شاخص چه معنایی دارد و چگونه محاسبه میشود.

پاسخ :

بله، علاوه بر شاخص آنتروپی، شاخص Gini impurity نیز برای ساخت درخت تصمیم استفاده می‌شود. شاخص Gini impurity یک معیار اندازه‌گیری است که درخت تصمیم را بر اساس تقسیم بندی بهینه داده‌ها ارزیابی می‌کند.
شاخص Gini impurity براساس احتمال اشتباه تقسیم داده‌ها در هر گره محاسبه می‌شود. این شاخص در محاسبه میزان "خلط" یا "پراکندگی" داده‌ها در هر گره نقش دارد. هرچقدر مقدار Gini impurity برای یک گره کمتر باشد، به این معنی است که داده‌ها در آن گره به طور خالصه‌تر و یکنواخت‌تر جمع شده‌اند.
برای ساخت درخت تصمیم، معمولاً از هر دو شاخص آنتروپی و Gini impurity استفاده می‌شود. هر دو شاخص می‌توانند به صورت کمینه کردن مقدار خلط داده‌ها در هر گره مورد استفاده قرار بگیرند. استفاده از یکی از این دو شاخص بستگی به مسئله مورد بررسی و ترجیح شما دارد.
تفاوت اصلی بین روش Gini Index و آنتروپی (Entropy) در مفهوم درخت تصمیم در محاسبه تفاوت (Impurity) می‌باشد.
1. آنتروپی: در محاسبه آنتروپی، از مفهوم اندازه‌گیری ترتیب یا نظم در داده‌ها استفاده می‌شود. آنتروپی یک معیار اطلاعاتی است که میزان بی‌نظمی و خلط داده‌ها را نشان می‌دهد. مقدار آنتروپی برای یک گره با توجه به توزیع کلاس‌ها در آن گره محاسبه می‌شود. هدف در این روش، کاهش آنتروپی و افزایش نظم و یکنواختی داده‌ها است. به عبارت دیگر، درخت تصمیم با استفاده از آنتروپی سعی می‌کند گره‌هایی را انتخاب کند که بیشترین اطلاعات را در مورد تقسیم بندی کلاس‌ها در خود دارند.
2. Gini Index: در محاسبه Gini Index نیز، میزان خلط و بی‌نظمی داده‌ها را اندازه‌گیری می‌کند، اما با استفاده از مفهوم احتمال انتخاب دو نمونه تصادفی از یک گره و تعیین اینکه آیا این دو نمونه به دو کلاس مختلف تعلق دارند یا خیر. بنابراین، Gini Index میزان خلط موجود در یک گره را نشان می‌دهد و هدف در این روش، کاهش Gini Index و افزایش تمیزی و جداپذیری کلاس‌ها در گره‌ها است.
به طور کلی، هر دو روش Gini Index و آنتروپی به منظور اندازه‌گیری خلط و بی‌نظمی داده‌ها و تقسیم بندی بهتر کلاس‌ها در گره‌های درخت تصمیم استفاده می‌شوند. تفاوت اصلی آنها در روش محاسبه آن ها است.
روش محاسبه Gini Index در درخت تصمیم برای اندازه‌گیری خلط و بی‌نظمی داده‌ها استفاده می‌شود. مقدار Gini Index برای یک گره بر اساس توزیع کلاس‌ها در آن گره محاسبه می‌شود.
فرمول محاسبه Gini Index برای یک گره با n کلاس به صورت زیر است:
(Gini Index = 1 - Σ(p_i^2
در این فرمول، p_i نسبت تعداد نمونه‌های کلاس i به کل نمونه‌های گره مورد نظر است. Σ(p_i^2) نیز مجموع مربعات نسبت‌های کلاس‌ها می‌باشد.

مقدار Gini Index بین 0 و 1 قرار می‌گیرد، که معنای آن به صورت زیر است:
- مقدار 0 به معنای این است که همه نمونه‌های گره به یک کلاس تعلق دارند و خلطی وجود ندارد.
- مقدار 1 به معنای این است که تمام کلاس‌ها با تراکم یکسان در گره توزیع شده‌اند و خلط بیشینه است.

بنابراین، هر چقدر مقدار Gini Index کمتر باشد، نظم و جداپذیری کلاس‌ها در گره بیشتر است و بهترین تقسیم بندی را نشان می‌دهد. انتخاب جداکننده‌هایی که باعث کاهش Gini Index می‌شوند، بهبود تقسیم بندی و جداپذیری کلاس‌ها را به ارمغان می‌آورد.

سوال یازدهم :

سوال یازدهم

درمورد مسائل رگرسیون به سوالات زیر پاسخ دهید :

الف) simple linear regression و multiple linear regression با یکدیگر مقایسه کرده و تفاوت و شباهت های آنها را بیان کنید.

ب) یکی از راه‌های جلوگیری از بیش برازش استفاده از منظم‌سازی[۷] است که به دو نوع L1 و L2 تقسیم میشود. به نوع اول Lasso Regression و به نوع دوم Ridge regression گفته میشود. تفاوت این دو روش را از نوع بهینه سازی بیان کرده و نحوه کار آنها را توضیح دهید.

ج) در جدول زیر سن و فشار خون چند بیمار قلبی داده شده است. معادله رگرسیون به فرم $y = \beta_0 + \beta_1 x$ به دست آورید. همچنین با استفاده از معادله به دست آمده فشار خون یک بیمار ۴۰ ساله را پیش بینی کنید.(متغیر x نشان دهنده سن و متغیر y نشان دهنده فشار خون است)

Patient	A	B	C	D	E	F	G
x	42	74	48	35	56	26	60
y	98	130	120	88	182	80	135

پاسخ :

الف :

در زمینه رگرسیون، تفاوت اصلی بین Simple Linear Regression و Multiple Linear Regression در تعداد متغیرهای وابسته است که در مدل استفاده می‌شود.

در Simple Linear Regression، یک متغیر وابسته و یک متغیر مستقل وجود دارد. این مدل به صورت یک خط راست برای پیش‌بینی متغیر وابسته بر اساس متغیر مستقل استفاده می‌شود. به عبارت دیگر، در Simple Linear Regression، رابطه بین یک متغیر وابسته و یک متغیر مستقل را مدلسازی می‌کنیم.

اما در Multiple Linear Regression، بیش از یک متغیر مستقل برای پیش‌بینی متغیر وابسته استفاده می‌شود. در این حالت، رابطه بین یک متغیر وابسته و چندین متغیر مستقل را مدلسازی می‌کنیم. به عبارت دیگر، Multiple Linear Regression امکان می‌دهد بیشترین تعداد متغیرهای مستقل ممکن را در مدل رگرسیون استفاده کنیم.

شباهت اصلی بین Simple Linear Regression و Multiple Linear Regression در استفاده از مدل خطی برای تخمین و پیش‌بینی متغیر وابسته است. هر دو روش از رابطه خطی برای تعامل بین متغیرهای وابسته و مستقل استفاده می‌کنند. همچنین، هر دو مدل در ارزیابی تأثیر متغیرهای مستقل بر متغیر وابسته مفید هستند.

به طور کلی، اگر تعداد متغیرهای مستقل بیشتر از یک متغیر باشد، ما باید از Multiple Linear Regression استفاده کنیم تا تأثیر همه این متغیرها را بر روی متغیر وابسته مدلسازی کنیم. اما اگر فقط یک متغیر مستقل داشته باشیم، Simple Linear Regression کافی خواهد بود.

ب:

Lasso regression یک روش رگرسیون خطی است که برای انتخاب متغیرها و کاهش اهمیت متغیرهای غیرضروری در مدل استفاده می‌شود. در واقع، Lasso regression یک روش مناسب برای انتخاب ویژگی (feature selection) در مدلسازی است.

در Lasso regression، همچنین به عنوان رگرسیون L1 معروف است، علاوه بر تعامل متغیرهای مستقل با متغیر وابسته، یک جمله جریمه (penalty term) به مدل اضافه می‌شود که شامل جمع مقادیر مطلق ضرایب متغیرهای مستقل است. این جمله جریمه باعث می‌شود که برخی از ضرایب برابر با صفر شوند، یعنی متغیرهایی که اهمیت کمتری در تخمین متغیر وابسته دارند حذف شوند.

استفاده از Lasso regression دارای چندین مزیت است:

1. انتخاب ویژگی (Lasso regression :(feature selection به خوبی متغیرهای غیرضروری را حذف می‌کند و تنها متغیرهای مهم را در مدل نگه می‌دارد. این باعث ساده‌تر شدن مدل و بهبود قابلیت تفسیر آن می‌شود.

2. مقاومت در برابر برهم‌کنش متغیرها: Lasso regression تمایل دارد در صورت وجود برهم‌کنش بین متغیرها، فقط یکی از آن‌ها را در مدل نگه دارد و دیگری را حذف کند. این باعث می‌شود مدل بهتر با مشکل برهم‌کنش متغیرها روبرو شود.

3. تنظیم پارامترها: با استفاده از جمله جریمه در Lasso regression، می‌توانیم میزان تأثیر ضرایب را کنترل کنیم.

Ridge regression یک روش رگرسیون خطی است که برای کاهش اهمیت متغیرهای غیرضروری و کنترل برازش زیاد (overfitting) در مدل استفاده می‌شود. این روش در واقع یک تغییر کوچک به روش Least Squares (رگرسیون خطی معمولی) اعمال می‌کند.

در Ridge regression، همچنین به عنوان رگرسیون L2 معروف است، به جمع مربعات ضرایب متغیرهای مستقل یک جمله جریمه (penalty term) اضافه می‌شود. این جمله جریمه باعث می‌شود که مقادیر ضرایب کوچکتر شوند و اهمیت متغیرهای غیرضروری کاهش یابد.

استفاده از Ridge regression دارای چندین مزیت است:

1. کاهش برازش زیاد: Ridge regression کمک می‌کند تا برازش زیاد مدل (overfitting) کاهش یابد. با افزایش جمله جریمه، مدل مجبور می‌شود به متغیرهای غیرضروری کمتر توجه کند و به همین ترتیب میزان برازش زیاد را کاهش می‌دهد.

2. مقاومت در برابر چگالش بیش‌ازحد: در صورت وجود متغیرهای کم اهمیت یا تعداد بیش‌ازحد متغیرها نسبت به نمونه‌ها، Ridge regression میزان چگالش بیش‌ازحد را بهبود می‌بخشد. این باعث کاهش تاثیرات نامناسب و اهمیت کمتر متغیرها در مدل می‌شود.

3. تنظیم پارامترها: مقدار پارامتر جمله جریمه را می‌توان تنظیم کرد تا میزان تأثیر ضرایب را کنترل کند. این به ما امکان می‌دهد تا تراز بین دقت مدل و تعداد متغیرها را تنظیم کنیم.

تفاوت اصلی بین Lasso regression و Ridge regression در جمله جریمه‌ای است که در هر یک از آن‌ها استفاده می‌شود و نحوه تاثیرگذاری آن بر ضرایب متغیرها.

در Lasso regression، از جمله جریمه L1 استفاده می‌شود که شامل جمع مقادیر مطلق ضرایب است. این باعث می‌شود که برخی از ضرایب برابر با صفر شوند، یعنی متغیرهایی که اهمیت کمتری در تخمین متغیر وابسته دارند حذف شوند. در نتیجه، Lasso regression منجر به انتخاب ویژگی (feature selection) می‌شود و متغیرهای غیرضروری حذف می‌شوند.

در Ridge regression، از جمله جریمه L2 استفاده می‌شود که شامل جمع مربعات ضرایب است. این جمله جریمه باعث کاهش اندازه ضرایب متغیرها می‌شود، اما به طور کلی ضرایبی که اهمیت بیشتری در تخمین متغیر وابسته دارند حذف نمی‌شوند. در نتیجه، Ridge regression منجر به کاهش برازش زیاد (overfitting) می‌شود و متغیرهای غیرضروری کمتر تأثیری در مدل دارند، اما حذف کامل آن‌ها رخ نمی‌دهد.

به طور کلی، تفاوت اصلی بین این دو روش در رویکرد انتخاب ویژگی است. Lasso regression تمایل دارد متغیرهای غیرضروری را حذف کند و تنها متغیرهای مهم را در مدل نگه دارد، در حالی که Ridge regression تمایل دارد ضرایب تمام متغیرها را کاهش دهد اما حذف کامل آن‌ها را نداشته باشد.

ج:

$B = (X^T X)^{-1} X^T y,\ x = \begin{pmatrix} 1 & 42 \\ 1 & 74 \\ 1 & 48 \\ 1 & 35 \\ 1 & 56 \\ 1 & 26 \\ 1 & 60 \end{pmatrix}$ y = $\begin{pmatrix} 98 \\ 130 \\ 120 \\ 88 \\ 182 \\ 80 \\ 135 \end{pmatrix}$ => $(X^T X) = \begin{pmatrix} 7 & 341 \\ 341 & 18181 \end{pmatrix}$

$(X^T X)^{-1} = \begin{pmatrix} \frac{18181}{10986} & -\frac{341}{10986} \\ -\frac{341}{10986} & \frac{7}{10986} \end{pmatrix}$ B= $\begin{pmatrix} \frac{499505}{10986} \\ \frac{16583}{10986} \end{pmatrix}$

$y = B_0 + B_1 x = 45.467413071 + 1.509466594x$

برای یک فرد ۴۰ ساله x را برابر ۴۰ قرار میدهیم

$y = 45.467413071 + 1.509466594 * 40 = 105.846076825$

بخش عملی:

رگرسیون:

رفتن از رگرسیون خطی به درجه ۲ نتایج را بهبود میدهد اما انجام رگرسیون درجه آنچنان تفاوت خاصی ایجاد نمی کند.

دسته بندی :

بین الگوریتم های پیاده سازی شده درخت تصمیم هم در داده های تست و هم در داده های آموزشی بالاترین دقت را داشته و بقیه معیار های آن که در کد وجود دارد نیز از بقیه بهتر هستند.